# UNPACKING HUMAN PREFERENCE FOR LLMS: DEMOGRAPHICALLY AWARE EVALUATION WITH THE DIVERSE FRAMEWORK

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The evaluation of large language models faces significant challenges. Technical benchmarks often lack real-world relevance, while existing human preference evaluations can be undermined by methodological limitations such as unrepresentative sampling, superficial assessment depth, and single-metric reductionism. To help address these issues, we introduce DIVERSE, a framework designed to provide a robust alternative by systematically addressing these key concerns. We collected multi-turn, naturalistic conversations from 21,352 participants, a sample stratified across 22 key demographic groups in the US and UK, to evaluate 27 state-of-the-art models across five human-centric dimensions. Using a hierarchical Bayesian Bradley-Terry-Davidson (BTD) model with post-stratification to census data, our analysis reveals several key insights. We establish **(1)** a clear performance hierarchy where `google/gemini-2.5-pro` ranks first overall, with our Bayesian analysis assigning it a 97% posterior probability of being the top-ranked model, indicating a high degree of statistical confidence in its lead. We uncover **(2)** significant preference heterogeneity, with user age emerging as the primary demographic axis of disagreement; a model's perceived rank can shift substantially across age groups, exposing failures in generalisation that unrepresentative samples typically mask. Finally, we quantify **(3)** the vast difference in discriminative power across evaluation dimensions, with ambiguous qualities like *Trust, Ethics & Safety* showing a 65% tie rate, in stark contrast to the decisive 10% tie rate for *Overall Winner*. Our work contributes a methodology and a set of findings that underscore the need for a more multidimensional, demographically aware perspective in LLM evaluation. We release our complete dataset, interactive leaderboard, and open-source framework to support the development of more rigorous and equitable evaluation practices.

## 1 INTRODUCTION

Large Language Models (LLMs) have facilitated a sea change in how humans interact with AI, becoming deeply integrated into professional workflows, personal decisions, and creative tasks. However, this rapid progress has created a critical "evaluation gap", our methods for measuring models have not kept pace with their real-world impact. This gap is perpetuated by the primary of automated benchmarks, which almost exclusively assess technical performance while overlooking how the systems resonate with the people who actually use them (Bowman & Dahl, 2021). As a result, optimising for benchmarks alone risks developing models that are technically impressive yet fail to meet human needs and expectations (Amershi et al., 2019), leaving the entire AI ecosystem without the reliable human-centric data needed to guide responsible development and deployment.

The field's dependence on automated benchmarks exemplifies this problem. Benchmarks like MMLU (Hendrycks et al., 2021), HELM (Liang et al., 2022), and BIG-Bench (Srivastava et al., 2022) are indispensable for establishing a model's technical floor by assessing its foundational reasoning and knowledge. However, their design as standardised tests makes them blind to the subjective, dynamic qualities of conversation. They fall short of measuring a model's ability to maintain context, adapt its tone, or build user trust. As the field orients around these metrics, development can fall prey to a form of "Goodhart's Law", where optimising for the benchmark becomes the goal,

rather than improving the holistic user experience the score was intended to represent. Ultimately, while these benchmarks measure what a model knows, they fail to capture how it behaves in the complex domain of human collaboration.

To address the limits of automated benchmarks, a second paradigm has emerged: direct human preference evaluation. Influential platforms like Chatbot Arena (Zheng et al., 2023) represent a crucial step forward by crowdsourcing pairwise comparisons from users in live conversations. However, their approach is undermined by foundational methodological flaws. First, their reliance on a self-selected, anonymous user base leads to unrepresentative sampling. Second, judgments often based on minimal interaction result in superficial assessment depth. Finally, binary preference votes create single-metric reductionism, obscuring the multidimensional nature of interaction quality. These inherent issues are compounded by systemic artefacts; as documented by Singh et al. (2025), practices like undisclosed private testing and evaluation gaming can distort rankings independently of true model quality.

To address these multifaceted issues, we introduce DIVERSE: a rigorous evaluation framework designed for multidimensional, demographically aware measurement of human-AI interaction. DIVERSE's methodology is built on the foundational principles of psychometric measurement and directly counters the flaws of existing approaches. To eliminate sampling bias, we employ demographically stratified sampling with post-stratification adjustments to census data. To ensure assessment depth, we mandate multi-turn conversations on participant selected topics. To move beyond single metric reductionism, we collect judgement across five distinct evaluation dimensions. The resulting framework provides a robust and scientifically grounded alternative for understanding the complex patterns that determine real-world model preference.

Our primary contributions are:

1. **The DIVERSE Framework:** A scientifically grounded methodology for human-centric AI evaluation that remedies the critical validity threats plaguing existing approaches: sampling bias, assessment depth, and metric reductionism.
2. **A Large-Scale, Demographically Stratified Dataset:** The release of a dataset to catalyse further research, containing (a) over 100,000 multi-dimensional human judgments on LLM performance and (b) structured metadata derived via an LLM judge, characterising the conversational dynamics, task properties, and interaction outcomes.
3. **Key Empirical Insights:** Our analysis provides (1) a clear performance hierarchy of 27 models, establishing `google/gemini-2.5-pro`'s top rank with a 97% posterior probability of being the top-ranked model; (2) identification of user age as the primary demographic driver of disagreement in model preference; and (3) quantification of the varying discriminative power across evaluation metrics.
4. **A Continuous Evaluation Framework:** The release of a dynamic evaluation platform featuring a continuously updated leaderboard and an evolving framework, designed to provide the community with an ongoing, current resource for tracking state-of-the-art model performance.

## 2 Related Works

The DIVERSE framework is situated at the intersection of several research domains: large language model (LLM) evaluation, psychometric measurement theory, human-computer interaction (HCI), and the study of fairness and representation in AI.

### 2.1 Paradigms in LLM Evaluation

Current LLM evaluation is dominated by two paradigms. The first, automated benchmarks, provides essential measures of technical capability through standardised tests like MMLU (Hendrycks et al., 2021) and HELM (Liang et al., 2022). DIVERSE complements this work by addressing their inability to capture subjective interaction quality (Bowman & Dahl, 2021). The second, human preference evaluation, pioneered by platforms like Chatbot Arena (Zheng et al., 2023), moved evaluation towards real-world interaction. Our framework is a direct response to the significant methodological

flaws of this approach, including unrepresentative sampling and systemic gaming artefacts (Singh et al., 2025).

A third emerging paradigm is model-based evaluation, or "LLM-as-a-judge" (Zheng et al., 2023; Liu et al., 2023). While this approach offers scalability, it is prone to a host of biases. DIVERSE instead adopts a complementary role for the LLM judge: not as a proxy for human preference, but as a tool for structured, post-hoc analysis of conversational content to help explain the phenomena underlying human judgments.

### 2.2 Psychometric Foundations for Preference Modelling

The conversion of pairwise comparisons into a continuous scale has a long history in psychometrics, originating with Thurstone's Law of Comparative Judgment (Thurstone, 1927) and formalized in models like the Bradley-Terry model (Bradley & Terry, 1952). The statistical engine of DIVERSE, a hierarchical Bayesian implementation of the Bradley-Terry-Davidson model, applies modern statistical techniques to this established measurement framework, allowing for robust uncertainty quantification and the modeling of complex, multi-level effects.

### 2.3 Human-Centric and Usability Frameworks

DIVERSE's multi-dimensional metrics are grounded in decades of research from Human-Computer Interaction (HCI). The framework aligns with concepts like the Technology Acceptance Model (TAM), which posits that "perceived usefulness" and "perceived ease of use" are primary determinants of technology adoption (Davis, 1989). Our dimensions map directly onto these ideas: "Core Task Performance" reflects usefulness, while "Interaction Fluidity" and "Communication Style" capture ease of use. This approach operationalises principles from the *Guidelines for Human-AI Interaction* (Amershi et al., 2019), which emphasise that AI systems must be understandable, adaptable, and trustworthy.

### 2.4 Representative Data and Fairness in AI Evaluation

A core contribution of DIVERSE is its commitment to representative sampling. The reliance on unrepresentative datasets has been shown to result in systems that perform inequitably across demographic groups, as famously demonstrated in facial recognition by Buolamwini & Gebru (2018). More recently, Santurkar et al. (2023) provided direct empirical evidence that rater demographics significantly impact preferences for LLM behaviour, showing that an aggregate score can mask important disagreements between populations. This aligns with findings from Kirk et al. (2024), who demonstrated the importance of demographically diverse preference data for safety alignment. By employing demographically stratified sampling with post-stratification adjustments, DIVERSE is explicitly designed to measure this preference heterogeneity, moving from assumption to quantification.

## 3 Methodology

### 3.1 Model & Participant Selection

For our first instantiation, we selected 27 state-of-the-art language models representing the current frontier of conversational AI, accessed via `openrouter.ai` with default settings. As DIVERSE is designed as a living benchmark, we continuously add new models and update rankings; the list of models is therefore a snapshot at the time of writing.

We recruited 21,352 participants through the Prolific platform, compensating them at the recommended rate of £9/hr. To enable deep demographic analysis, we stratified our sampling to include 22 specific demographic strata across geographic location (UK/US), age (18-34, 35-54, 55+), ethnicity (Asian, Black, White and Other in the UK, and Hispanic, Asian, African American, and White in the US.), and political affiliation (Democrat, Republican, and Independent in the US, and Conservative, Labour, Liberal Democrats, Greens, and Reform UK in the UK). For more detail on the participant experience, refer to Appendix B.

### 3.2 Data Collection and Benchmark Design

The DIVERSE benchmark employs a pairwise comparison framework. Participants were presented with two anonymised models side-by-side and were free to select their own conversation topic. To ensure sufficient interaction depth, a minimum of 3 conversational turns was required. Each message sent by the participant was delivered to both models simultaneously, ensuring identical user-side conversation flow for a controlled comparison.

To maximise data collection efficiency, model pairings were determined by a TrueSkill-based adaptive sampling algorithm (Herbrich et al., 2006). By maintaining skill and uncertainty estimates for each model, the algorithm strategically selects matchups where the outcome is most uncertain, thereby maximising information gain and accelerating the convergence of rankings.

Each of our 22 demographic strata was run as a dedicated TrueSkill tournament. Participants could qualify for and participate in multiple tournaments corresponding to their demographic profile (e.g., a Hispanic, 18-34 Democrat could participate in three separate tournaments). Additionally, participants could take part in multiple data collection batches, receiving new, randomly assigned model pairs each time. The statistical implications of this multi-membership design are handled by our hierarchical model, as detailed in Subsection 3.4.1.

Finally, conversation quality was monitored in real-time by a `gpt-4o-mini` judge that flagged low-effort inputs, providing participants with constructive feedback. Participants receiving three warnings were removed from the study. After the conversation, participants evaluated the two models across the five comparative dimensions, selecting their preferred model or indicating a tie.

More details on data collection and the quality assurance mechanism are available in Appendix B.

### 3.3 Evaluation Metrics

Our four evaluation dimensions were derived from a pilot study using factor analysis to identify the core drivers of user preference. To these four dimensions, we added a holistic overall winner metric:

- **Core Task Performance & Reasoning:** How effectively the model accomplishes tasks and demonstrates sound reasoning and understanding.
- **Communication Style & Presentation:** The model's language, tone, personality, and the appropriateness of its detail and intuitiveness.
- **Interaction Fluidity & Adaptiveness:** How smoothly and adaptively the model interacts, manages conversation flow, and responds to user input.
- **Trust, Ethics & Safety:** The reliability, transparency, ethical conduct, and safety of the model's outputs and behaviour.
- **Overall Winner:** A holistic preference judgement incorporating all aspects of the interaction.

### 3.4 Analysis Framework

#### 3.4.1 Hierarchical Bradley-Terry-Davidson Model

We employ a hierarchical Bayesian Bradley-Terry-Davidson (BTD) model to convert pairwise comparisons into continuous skill ratings. The model extends the classic BT framework to handle ties and captures demographic heterogeneity through a factorised structure. At its core, it learns a global skill parameter ($\theta$) for each model-metric combination, then adds demographic-specific adjustments ($u$). These adjustments are hierarchically modelled with heterogeneity parameters ($\tau$) that quantify the magnitude of preference variation. The model outputs posterior distributions for all parameters, enabling us to derive global leaderboards, demographic-specific rankings, and measures of preference heterogeneity.

**Disentangling Mixed Demographic Effects with Hierarchical Modelling.** A key challenge of our participant design is that a single participant's preference (e.g., from someone who is Asian, 18-34, and a Democrat) could be driven by any of their demographic identities. Our single, unified hierarchical model is designed to disentangle these mixed effects. The tournament structure is purely a data-collection device; for the analysis, all comparisons are pooled.

The model represents each participant by their position on three demographic axes (age, ethnicity, politics). The model then learns two components simultaneously through partial pooling: a global skill parameter for each model, and a set of additive adjustments for each demographic group. These group adjustments are centred within each axis, allowing them to be interpreted as deviations from the average preference for that axis. Critically, this design allows the model to attribute a consistent preference pattern to the correct demographic driver, even when a participant belongs to several groups, while leveraging the entire dataset to ensure the global skill estimates remain robust.

Full mathematical details are provided in Appendix A.

#### 3.4.2 LLM JUDGE FOR CONVERSATIONAL ANALYSIS

To provide a deeper, quantitative understanding of the conversations underlying human preferences, we conducted a post-hoc analysis of all conversation transcripts using an LLM judge.

**Model Selection and Justification.** For this role, we selected `gpt-4.1`, balancing three key considerations: performance, practicality, and precedent. The model offered a strong trade-off between state-of-the-art instruction-following and the inference speed needed to process our large dataset. Its availability via a stable API was critical for reproducibility, and its capabilities in similar annotation tasks have been demonstrated in prior literature and validated by our internal testing.

**Procedural Safeguards.** While acknowledging that no LLM-based analysis can be entirely free of bias, we implemented several procedural safeguards to ensure the integrity and utility of the outputs. The core principle of our approach was strict separation: the LLM analysis was conducted entirely post-hoc, was never used to generate competitive rankings, and had no influence on the primary human preference scores. Its role was purely explanatory. To enhance reliability, we used a detailed, structured prompt with explicit rubrics. This systematic approach allows us to leverage the scalability of LLMs to generate rich metadata that characterises conversational dynamics, task properties, and outcomes as an explanatory tool for understanding human judgments. Full details on the metrics, categories, and prompt are provided in Appendix C.

## 4 RESULTS

Our analysis is based on 106,760 pairwise comparisons from 21,352 participants across 27 language models. We structure our findings into four parts: (1) we establish the overall model performance leaderboard; (2) we quantify the significant heterogeneity in preferences across demographic groups; (3) we examine how model rankings vary by evaluation dimension; and (4) we assess the discriminative power of each metric.

### 4.1 OVERALL MODEL PERFORMANCE

Our primary result is a robust ranking of the 27 models derived from our hierarchical BTD model, post-stratified to US and UK census data. The main performance metric, shown in Figure 1, is the Score (Winshare). This score represents a model's expected total points from a round-robin tournament against all other models, where a win is worth 1 point and a tie is worth 0.5, for a maximum possible score of 26.

`google/gemini-2.5-pro` stands out as the clear winner. Its leading position is substantiated not only by its top score of 18.2, but also by the high statistical confidence assigned to its rank: our Bayesian model calculates a 97.2% probability of it being the best model (P(best)). A distinct gap separates it from the next competitor, `deepseek/deepseek-chat-v3-0324` (score: 17.3), which in turn holds a lead over the subsequent tier of models.

Below the top two, a competitive group including `mistralai/magistral-medium-2026`, `x-ai/grok-4`, and `x-ai/grok-3` emerges with closely overlapping credible intervals. This establishes a clear hierarchy at the top of the leaderboard that progressively flattens, rendering many lower-ranked models statistically indistinguishable.

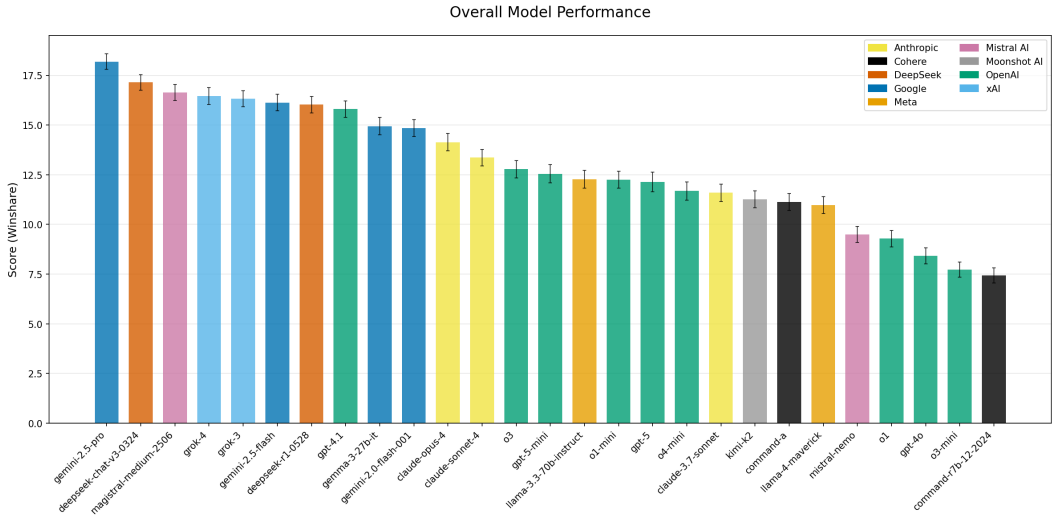



Figure 1: Model performance on the "Overall Winner" metric. Bars represent the Score (expected points in a round-robin tournament; max=26, mean=13), with 95% credible intervals.

### 4.2 Demographic Heterogeneity in Model Preference

A key objective of the DIVERSE framework is to move beyond a single aggregate leaderboard to quantify how model preferences vary across different populations. Our analysis reveals that **age is the most significant demographic factor driving this preference heterogeneity**, substantially exceeding the effects of ethnicity and political affiliation. While our hierarchical model quantifies this through a latent heterogeneity parameter ($\tau$), the practical impact of this finding is best understood through the two more interpretable metrics visualised in Figure 2.

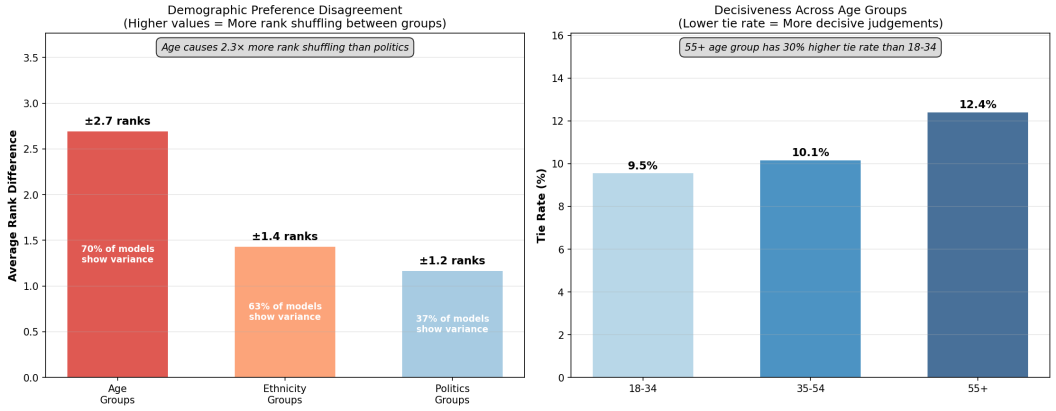



Figure 2: Demographic preference heterogeneity, shown by: **(Left)** inter-group disagreement (avg. rank difference), and **(Right)** user decisiveness (tie rates by age).

The **left panel** shows that a model's average rank shifts by a substantial ±2.7 ranks across age cohorts, a far larger variance than for ethnicity (±1.4) or political affiliation (±1.2). The **right panel** reveals a clear trend in user decisiveness: tie rates increase steadily with age, from 9.5% for the 18-34 cohort to 12.4% for users aged 55+, representing a 30% rise in indecisiveness. Together, these findings empirically validate our central claim that a single aggregate leaderboard is insufficient, as it masks critical variations in both inter-group agreement and user decisiveness.

### 4.3 Performance Across Evaluation Dimensions

While the overall leaderboard provides a valuable summary, it obscures critical nuances in model performance. The rankings across all five evaluation dimensions reveal that a model's competitive standing can change dramatically depending on the evaluation lens.

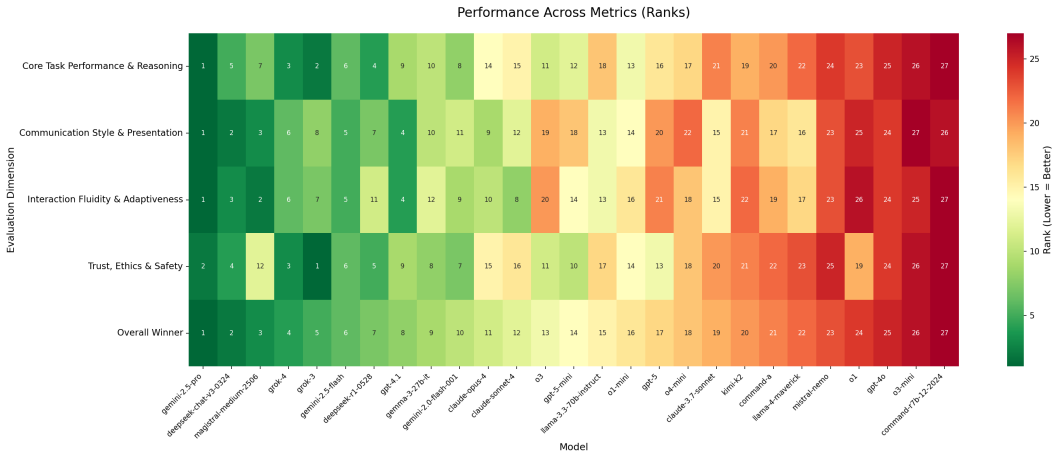
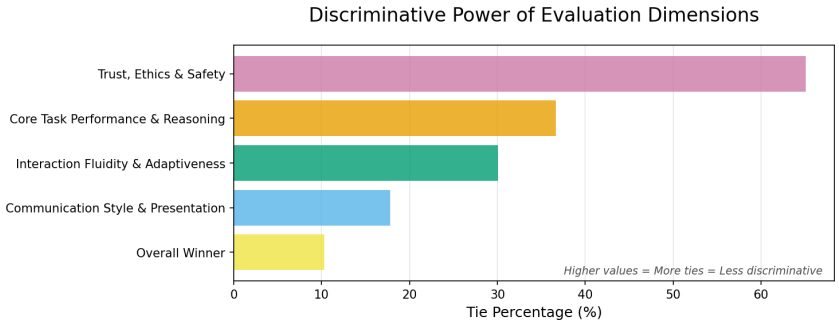



Figure 3: Heatmap showing model rankings across five evaluation dimensions. Lower ranks (darker green) indicate better performance. Models show significant variation in their relative strengths, with some excelling in reasoning while others lead in communication or trust.

While `google/gemini-2.5-pro` maintains the top position across all dimensions, the rankings of other models shift significantly. Notably, `x-ai/grok-3` performs substantially better on *Core Task Performance & Reasoning* (ranking 2nd) than on *Communication Style & Presentation* (ranking 8th). Conversely, `mistralai/magistral-medium-2506` excels in communication (ranking 3rd) but ranks lower in task performance (ranking 7th). These shifts underscore the multi-faceted nature of human preference and highlight the danger of relying on a single "overall" score for model selection.

### 4.4 Discriminative Power of Evaluation Dimensions

The BTD model's tie-propensity parameter ($\nu_k$) allows us to assess the discriminative power of each evaluation dimension. We observe substantial variation in how decisively participants can distinguish between models across different metrics.




*Trust, Ethics & Safety* was the least discriminative dimension with a 65% tie rate, suggesting either model convergence on safety or that the quality is inherently difficult to assess in brief interactions. In stark contrast, *Overall Winner* was the most discriminative (10% ties), indicating that users can form decisive holistic preferences even when specific attributes are ambiguous.

This metric hierarchy suggests pairwise comparison's effectiveness is highly dependent on the evaluation metric. Holistic judgments like *Overall Winner* provide a strong, decisive signal in open-ended

conversations. Conversely, the high ambiguity of *Trust, Ethics & Safety* indicates a more tailored approach is needed to elicit relevant behaviours.

## 5 DISCUSSION

The DIVERSE framework advances the evaluation of large language models by moving beyond single-metric leaderboards to reveal the multidimensional and demographically-contingent nature of human-AI interaction quality. Our analysis uncovered three critical insights that challenge current evaluation paradigms and offer a new path forward for model development, assessment, and selection.

First, **dimensional analysis reveals that "best" is a context-dependent illusion.** While confirming that `google/gemini-2.5-pro` leads across multiple dimensions, DIVERSE demonstrates *why* it succeeds: through balanced, high performance. Our findings show models exhibit different competitive standings depending on the evaluation lens: `x-ai/grok-3`, for example, leads on *Trust, Ethics & Safety* but ranks 8th on *Communication Style & Presentation*, while `mistralai/magistral-medium-2506` excels in communication (3rd) but lags in task performance (7th). This multifaceted performance becomes particularly striking when contrasted with technical benchmarks like HELM, where `google/gemini-2.5-pro` currently ranks a modest 13th, a dramatic disparity highlighting the evaluation gap between technical accuracy and human preference. These insights disappear when collapsed into a single number, leaving users unable to discern if a model's success is due to its reasoning power, communication skills, or balanced competence. DIVERSE moves the conversation from *"which model is best?"* to *"best for what and for whom?"* and demonstrates that meaningful model selection requires aligning specific dimensional strengths with intended use cases.

Second, **the discovery that age is the primary driver of preference heterogeneity exposes a critical demographic blind spot in AI development.** Current evaluation practices that rely on unrepresentative user bases systematically obscure these crucial performance gaps. Our analysis reveals that user preferences shift by ±2.7 ranks across age cohorts, far exceeding variation for ethnicity (±1.4) or politics (±1.2). Strikingly, the pattern of tie rates reveals how age shapes evaluation certainty: while all age groups show similar decisiveness about *Communication Style & Presentation* (17-20% ties), older users become progressively less certain about *Core Task Performance & Reasoning* (rising from 32% ties for 18-34 to 39% for 55+). This suggests younger users have clearer expectations for functional capability, while older users find it harder to distinguish between models on core utility, potentially reflecting different mental models of what AI should accomplish. The overall trend toward higher tie rates among older users (from 9.5% for 18-34 to 12.4% for 55+) indicates that qualities differentiating models for younger demographics are systematically less salient for other groups. These findings suggest that models tuned on narrow, tech-savvy feedback risk creating preference-optimisation loops that systematically exclude broader populations, undermining both market adoption and equitable performance.

Third, **the vast difference in metric discriminability reveals that evaluation methodologies must be tailored to the constructs they aim to measure.** Our analysis shows that the context of an interaction is critical for reliably assessing certain qualities. The 65% tie rate for *Trust, Ethics & Safety* suggests that these qualities are not consistently elicited in open-ended, user-driven conversations, making them difficult for participants to meaningfully compare. This finding offers a clear methodological principle for the field: while broad-based preference testing like DIVERSE is highly effective for measuring general utility, assessing critical but nuanced attributes like perceived safety demands a move towards more specialised interaction scenarios that create the necessary context for meaningful judgment.

### 5.1 LIMITATIONS AND FUTURE WORK

While our methodology addresses key flaws in existing paradigms, we acknowledge several limitations that provide avenues for future research. Our initial study is confined to US and UK participants, limiting global applicability as cultural context can profoundly influence preferences. Our demographic stratification could also be extended to include other factors like gender, education, and socioeconomic status, which may reveal additional layers of preference heterogeneity.

Our focus on short, multi-turn conversations cannot capture long-term phenomena like persona consistency or performance degradation over extended dialogues. Moreover, the open-ended nature of the tasks, while ecologically valid, means task complexity was not controlled. Future work should incorporate longer-term interactions and controlled, multi-step tasks.

The five evaluation dimensions, while empirically derived, may not be exhaustive. Qualities like creativity, humour, or empathy may be significant preference drivers in certain contexts. Additionally, the importance of these dimensions may vary across cultures, reinforcing the need for a globally-expanded framework.

The DIVERSE framework is currently limited to text-only interactions. This represents a growing gap, as the state-of-the-art models under evaluation are increasingly capable of processing and generating images, audio, and other data types. A text-only evaluation, therefore, assesses only a fraction of their true capabilities and utility in real-world use cases. Future iterations of the framework must incorporate multimodal interactions, presenting a significant research challenge in designing tasks that evaluate not just the quality of outputs in each modality, but also the model's ability to reason and converse coherently across them.

Furthermore, our open-ended conversational design proved to be an imprecise tool for assessing specific, nuanced qualities. As discussed, the high tie rate for *Trust, Ethics & Safety* indicates this methodology does not reliably create a context where such judgments can be made. This limitation points to a clear direction for future research: developing targeted evaluation suites that measure subjective preferences within specialised scenarios. For instance, future studies could use a pairwise comparison framework to evaluate how different models handle sensitive topics, navigate ethical boundaries, or respond to requests for advice in high-stakes domains. Such a focused approach would provide the necessary context for users to form discriminative judgments, yielding a much stronger signal than is possible with generic interactions.

## 6 Conclusion

The evaluation of large language models requires moving beyond the pursuit of a single, universal score. The DIVERSE framework offers a methodology for this shift, demonstrating that an over-reliance on aggregate scores is insufficient because it obscures critical performance trade-offs, masks demographic blind spots, and misrepresents the utility of different evaluation metrics.

These findings underscore the need for a more nuanced approach to AI development and deployment. For developers, our results highlight the challenge of navigating performance trade-offs across diverse users, rather than simply optimising a singular metric. For organisations, it points towards the importance of a context-aware selection process that aligns a model's specific strengths with their users' needs.

To support this effort, we release our dataset, leaderboard, and framework as public resources. Critically, DIVERSE is designed as a live benchmark: the leaderboard is continuously updated with new models and fresh human evaluations, ensuring it remains a current reflection of the state-of-the-art. This continuous evaluation approach, which prizes nuance over numbers, is a foundational step towards an evaluation practice that helps catalyse research into AI that is demonstrably equitable, reliable, and genuinely beneficial for the diverse human populations it is meant to serve.

## References

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–13, 2019.

Samuel R Bowman and George E Dahl. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. *arXiv preprint arXiv:2108.12721*, 2021.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91, 2018.

Fred D Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pp. 319–340, 1989.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2021.

Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill: A bayesian skill rating system. In *Advances in neural information processing systems*, pp. 569–576, 2006.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Prism: A preference-based safety alignment dataset for large language models. *arXiv preprint arXiv:2406.01315*, 2024.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023.

Amit Singh et al. The leaderboard illusion: Biases and gaming in human preference benchmarking. *arXiv preprint arXiv:2504.20879*, 2025.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Louis L Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

# A STATISTICAL METHODOLOGY: HIERARCHICAL BRADLEY-TERRY-DAVIDSON MODEL

This section formalises the model that turns human A/B/Tie judgements into the leaderboard statistics. It uses a **Bradley–Terry–Davidson (BTD)** outcome model with **hierarchical** demographic adjustments and **post-stratification** to census weights.

## A.1 THE OUTCOME MODEL: PREDICTING A CHOICE

At its core, the model predicts the outcome of a single comparison based on the "latent advantage" ($\eta$) of model $i$ over model $j$. This advantage is the sum of the difference in their baseline skills and the difference in their demographic effects for that specific rater.

The demographic effect for a model ($\Delta u$) is the sum of its adjustments across the rater's age, ethnicity, and political groups:

$$\Delta u_{i,\text{rater}} = u^{\text{age}}_{i,g_a,k} + u^{\text{eth}}_{i,g_e,k} + u^{\text{pol}}_{i,g_p,k} \tag{1}$$

If a rater has multiple labels on an axis, we treat that axis as **multi-membership** by taking the equal-weight average of the corresponding group adjustments (the weights on that axis sum to 1). If an axis is unobserved for a rater, its contribution is set to 0.

The total advantage, $\eta$, is then:

$$\eta = \underbrace{(\theta_{i,k} - \theta_{j,k})}_{\text{Baseline Skill Difference}} + \alpha \underbrace{(\Delta u_{i,\text{rater}} - \Delta u_{j,\text{rater}})}_{\text{Demographic Effect Difference}} \tag{2}$$

We scale demographic effects by $1/\sqrt{3}$ so that the combined effect of three demographic axes remains on the same scale as a single axis.

Given this advantage $\eta$, the probabilities for each outcome (A wins, Tie, B wins) are calculated using the BTD formula, which includes a per-metric **tie propensity** $\nu_k > 0$:

$$p_A = \frac{e^{\eta}}{Z}, \qquad p_T = \frac{\nu_k}{Z}, \qquad p_B = \frac{e^{-\eta}}{Z} \qquad \text{where } Z = e^{\eta} + e^{-\eta} + \nu_k \tag{3}$$

## A.2 PRIORS AND LATENT STRUCTURE: HOW PARAMETERS ARE LEARNED

The model's parameters are learned from the data using the following structure and priors:

- **Baseline Skill ($\theta$):** To ensure the skills are identifiable, we enforce a zero-sum constraint for each metric $k$: $\sum_i \theta_{i,k} = 0$
- **Demographic Adjustments ($u$):** The adjustments are learned **hierarchically** to ensure stability (a technique called partial pooling). For each demographic axis (e.g., age), the adjustments are centred and scaled by a parameter $\tau$ which is learned from the data.

  $$u^a_{i,y,k} = \left(u^a_{\text{raw},i,y,k} - \overline{u^a_{\text{raw},i,\cdot,k}}\right) \tau^a_k \tag{4}$$

  The raw, unscaled adjustments are drawn from a standard normal distribution, $u_{\text{raw}} \sim N(0,1)$, and the scale parameter $\tau$ (the "volume knob") is drawn from an exponential distribution, $\tau \sim \text{Exponential}(\lambda = 12)$.

## A.3 POPULATION ADJUSTMENT: REFLECTING THE REAL WORLD

After learning the parameters from our participants, we create a **population-adjusted skill** for each model by taking the expectation of the demographic effects, weighted by census data ($w$). For each posterior draw, this is:

$$\theta^{\text{pop}}_{i,k} = \theta_{i,k} + \alpha\Big(\langle w_{\text{age}}, u^{\text{age}}_{i,\cdot,k}\rangle + \langle w_{\text{eth}}, u^{\text{eth}}_{i,\cdot,k}\rangle + \langle w_{\text{pol}}, u^{\text{pol}}_{i,\cdot,k}\rangle\Big) \tag{5}$$

Here, $\langle w, u\rangle$ represents the dot product (a weighted average) of the census weights with the model's demographic adjustments for that axis.

### A.4 Scoring and Leaderboard Construction

From the population-adjusted skills, we construct the final leaderboard metrics for each posterior draw:

- The **Expected Points (Winshare)** for model $i$ vs. $j$ is: $\mathrm{EP}_{i \text{ vs } j,k} = p_A + \frac{1}{2}p_T$
- A model's **Score** for that draw is the sum of its EP against all opponents.
- Aggregating these Scores across all posterior draws gives us the final **mean Score**, its uncertainty interval, the **Expected Rank**, and the **P(best)**.

## B Detailed Methodology

This section provides additional details on the participant experience, data collection procedures, and quality assurance mechanisms employed in the DIVERSE framework.

### B.1 Participant Experience and Interface

Participants accessed the evaluation interface through a web-based platform that presented two AI models side-by-side in an anonymised format (labeled as "Model A" and "Model B"). The interface was designed to minimise cognitive load while ensuring thorough evaluation:

- **Topic Selection:** Participants began by choosing their own conversation topic, ensuring natural engagement and leveraging their domain expertise
- **Synchronised Input:** A single input field sent identical messages to both models simultaneously, ensuring perfect experimental control
- **Real-time Responses:** Models responded in parallel with streaming text, allowing participants to observe differences in response speed and style
- **Turn Requirements:** A minimum of 3 conversational turns was enforced before evaluation options became available
- **Evaluation Interface:** After conversation completion, participants rated models across five dimensions using a three-option format (Model A better, Tie, Model B better)

### B.2 Quality Assurance Framework

Our multi-layered quality assurance system balanced data quality with participant experience:

#### B.2.1 Real-time AI Monitoring

An AI evaluator (`gpt-4o-mini`) analysed messages in real-time to detect:

- Low-effort responses (e.g., single words, repetitive patterns)
- Disjointed conversation flow (unrelated topic jumping)
- Gaming behaviour (attempts to manipulate the system)

When issues were detected, participants received immediate, constructive feedback encouraging higher-quality engagement. The system used a warning-based approach, providing participants opportunities to improve rather than immediate exclusion.

#### B.2.2 Conversation Consistency Verification

To ensure fair model comparison, the system enforced that identical messages were sent to both models. Any attempt to send different messages to each model resulted in immediate task termination, as this would compromise the validity of the comparison.

### B.3 Compensation and Participation Structure

- **Fair Compensation:** All participants were compensated at £9 per hour, meeting ethical standards for research participation

- **Multi-demographic Participation:** Participants qualifying for multiple demographic groups could complete the task once for each relevant group, receiving full compensation for each completion
- **Batch Participation:** Across multiple data collection batches, returning participants received new, randomly assigned model pairs to prevent learning effects
- **Time Investment:** The median task completion time was approximately 15 minutes, including conversation and evaluation phases

### B.4 Continuous Framework Evolution

The DIVERSE framework is designed for continuous operation rather than one-time evaluation. Key aspects of our ongoing approach include:

- **Regular Model Updates:** New models are added to the evaluation pool as they become available, with the leaderboard updated monthly
- **Temporal Tracking:** Model performance is tracked over time to identify improvements or regressions across versions
- **Adaptive Sampling:** The TrueSkill-based algorithm continuously optimises model matchups based on uncertainty, focusing data collection where it provides maximum information gain
- **Expanding Demographics:** The framework is designed to incorporate additional demographic dimensions and geographic regions as the platform scales

This continuous evaluation approach ensures the leaderboard remains a living resource that reflects the current state of model capabilities rather than a static snapshot, providing ongoing value to researchers and practitioners.

## C LLM Judge Implementation Details

### C.1 Metrics and Classifications

To provide a comprehensive and structured characterisation of each conversation, our LLM judge was prompted to generate outputs across three distinct categories: quantitative metrics, categorical classifications, and a detailed qualitative analysis. This multi-faceted approach provided a rich layer of metadata for explaining the patterns observed in human preference data.

**Quantitative Metrics**. The judge assessed each conversation on a 1-5 scale across four independent axes to capture different aspects of the interaction quality and dynamics:

- **Task Complexity:** Measured the cognitive demand of the user's task, ranging from simple fact retrieval (1) to expert-level creative or abstract problem-solving (5).
- **Goal Achievement:** Assessed the degree to which the user's primary objective was accomplished, from complete failure (1) to the model exceeding expectations by providing proactive value (5).
- **User Satisfaction:** Inferred the user's sentiment from their language, ranging from explicit frustration (1) to enthusiastic praise or delight (5).
- **User Engagement:** Quantified the depth of the user's involvement, from a single transactional turn (1) to a deep, collaborative process over an extended dialogue (5).

**Categorical Classifications**. In addition to the quantitative scores, the LLM judge was tasked with classifying each conversation to provide a high-level understanding of its nature.

- **Task Type:** The primary activity the user was engaged in, chosen from a predefined list of 17 types (e.g., *information seeking*, *creative writing*, *coding & debugging*).
- **Domain:** The main subject area of the conversation, chosen from a list of 20 domains (e.g., *technology*, *health & medical*, *finance*).

These classifications allow for large-scale analysis of the types of tasks users naturally bring to LLMs and how preferences may vary across different domains.

### C.2 Analysis Prompt

This section provides the full prompt used for our conversation analysis.

```
You are an expert conversation analyst. Your goal is to score and
categorize a conversation between a user and an AI assistant with
high fidelity, using the entire 1-5 scale for metrics to differentiate
performance. Avoid clustering scores in the middle.

First, you will perform a step-by-step analysis. In a <reasoning>
block, you will provide a brief justification for each metric score,
explicitly referencing the scoring criteria.

Second, after your reasoning, you will provide the final output in
the required JSON format.

CONVERSATION:
{conversation_content}

--------------------------------

ANALYSIS TASK:

**Step 1: Provide your reasoning within <reasoning> tags.**
For each metric, briefly explain WHY you are choosing a specific
score, referencing the criteria below.

<reasoning>
- **Task Complexity Rationale**: [Explain why the user's task
  deserves a score of 1, 2, 3, 4, or 5 based on the cognitive
  demand. Reference the user's specific requests.]
- **Goal Achievement Rationale**: [Explain the extent to which
  the user's goal was met, partially met, or not met at all.
  Point to evidence in the text.]
- **User Satisfaction Rationale**: [Analyze the user's sentiment.
  Is there explicit praise, frustration, or just neutral acceptance?
  Quote or reference user language.]
- **User Engagement Rationale**: [Describe the depth of the
  interaction. Was it a simple transaction or a deep, collaborative
  exploration? Justify your score.]
</reasoning>

**Step 2: Based on your reasoning above, provide the final JSON.**
Return your analysis in this EXACT JSON format, with no other text
outside the JSON block.

'''json
{
  "metrics": {
    "task_complexity": 0,
    "goal_achievement": 0,
    "user_satisfaction": 0,
    "user_engagement": 0
  },
  "categories": {
    "task_type": "information_seeking",
    "domain": "religion",
    "complexity_tier": "medium",
    "engagement_tier": "moderate"
```

```
  },
  "detailed_analysis": {
    "conversation_starter": "direct_question",
    "user_initiative": "high",
    "model_proactiveness": "appropriate",
    "goal_achievement_evidence": "User got answer but asked follow-up",
    "primary_user_goal": "Learn about religious practices"
  }
}
```

**CRITICAL SCORING GUIDANCE (for "metrics"):**
- **Use the FULL 1-5 Scale**: You MUST use scores of 1, 2, 4, and 5 when warranted. A score of 3 is for truly average cases, not a default. If a task is a simple fact lookup, it is a 1. Do not inflate it.
- **Be a Strict Grader**: Your goal is to create distinctions. Scrutinize for flaws and unmet needs.

**SCORING CRITERIA (1-5 Scale, for "metrics"):**

**TASK_COMPLEXITY** - Cognitive demand on the user and model.
1: Simple fact retrieval. Single, unambiguous question.
   E.g., "What is the capital of France?"
2: Simple procedure or explanation. E.g., "How do I boil an egg?"
3: Requires synthesis of a few ideas or multi-step reasoning.
   E.g., "Compare Python and Java for web development."
4: Complex problem-solving or creating a nuanced argument.
   E.g., "Debug this complex code with a race condition."
5: Expert-level, creative, or highly abstract task.
   E.g., "Develop a market entry strategy for South America."

**GOAL_ACHIEVEMENT** - Was the user's objective accomplished?
1: Goal completely failed. User is explicit about failure or abandons.
2: Goal mostly failed. Core need is unmet.
3: Goal partially met. Main question answered, but user needs follow-up.
4: Goal fully met. User's stated goal is clearly accomplished.
5: Goal exceeded. Model was proactive and provided value beyond request.

**USER_SATISFACTION** - User's sentiment about the interaction.
1: Explicit frustration or anger.
2: Implicit frustration, impatience, or mild disappointment.
3: Neutral. Purely transactional.
4: Positive. User says "thanks," "perfect," or other positive indicators.
5: Enthusiastic. User expresses strong praise or delight.

**USER_ENGAGEMENT** - Depth of user's active involvement.
1: Single turn. One question, one answer.
2: Minimal follow-up. One or two simple clarifying questions.
3: Moderate exploration. User asks several related questions.
4: Active collaboration. User refines prompts, challenges the model.
5: Deep co-creation. Extended dialogue building complex understanding.

**CATEGORIES (for "categories"):**

**TASK_TYPE**: Choose from: information_seeking, technical_assistance, creative_writing, problem_solving, tutoring_explanation, brainstorming, research, writing_help, coding_debugging, analysis_review, project_planning, personal_advice, social_conversation, content_creation,

```
learning_education, decision_support, comparison_analysis.

**DOMAIN**: Choose from: technology, science, business, education,
health_medical, creative_arts, finance, law_legal, cooking_food,
travel, relationships, career_professional, academic_research,
programming, design, entertainment, sports, religion_philosophy,
history, language_linguistics.

**TIER CLASSIFICATIONS**:
- **COMPLEXITY_TIER**: low (scores 1-2), medium (3), high (4-5)
- **ENGAGEMENT_TIER**: low (scores 1-2), moderate (3), high (4-5)
```

### C.3 Implementation Notes

Conversations exceeding 50 turns were truncated to prevent context window issues. The prompt employs a two-stage reasoning approach where the LLM must first articulate its reasoning for each score before generating the final JSON output. This design reduces anchoring bias, the tendency for the model to commit to an initial score and then rationalise it post-hoc, by ensuring scores are grounded in explicit reasoning rather than intuition. Failed API calls or JSON parsing errors were logged and excluded from analysis. Importantly, all analysis was conducted post-hoc with no influence on human preference rankings.